PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68

A2

(11) International Publication Number: **WO 00/00637**

(43) International Publication Date: 6 January 2000 (06.01.00)

(21) International Application Number: PCT/CA99/00589

(22) International Filing Date: 25 June 1999 (25.06.99)

(30) Priority Data:
60/090,887 26 June 1998 (26.06.98) US

(71) Applicant: VISIBLE GENETICS INC. [CA/CA]; 700 Bay Street, Box 333, Toronto, Ontario M5G 1Z6 (CA).

(72) Inventors: GILCHRIST, Rodney, D.; 1114 Edgehill Place, Oakville, Ontario L6M 2E7 (CA). DUNN, James, M.; 117 Citadel Drive, Scarborough, Ontario M1K 4S8 (CA).

(74) Agent: DEETH WILLIAMS WALL; National Bank Building, Suite 400, 150 York Street, Toronto, Ontario M5H 3S5 (CA).

(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: METHOD FOR SEQUENCING NUCLEIC ACIDS WITH REDUCED ERRORS

(57) Abstract

In accordance with the invention, nucleic acid polymers are sequenced in a method comprising the steps of: a) obtaining forward and reverse data sets for forward and reverse strands of the sample nucleic acid; b) determining the apparent sequence of bases for the forward and reverse data sets; c) comparing the apparent forward and reverse sequences of basis for perfect complementarity to identify any deviations from complementarity in the apparent sequence, any such deviation presenting a choice between two bases, only one of which is correct; d) applying a confidence algorithm to peaks in the data set associated with a deviation to arrive at a numerical confidence value; and e) comparing each numerical confidence value to a predetermined threshold and selecting as the correct base the base represented by the peak which has the better numerical confidence value, provided that the numerical confidence value is better than the threshold. The confidence algorithm takes into account at least one, and preferably more than one of several specific characteristics of the peaks in the data sets that were not complementary.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

METHOD FOR SEQUENCING NUCLEIC ACIDS WITH REDUCED ERRORS

Background of the Invention

During routine sequencing of DNA from samples (such as HIV genotyping after RT-PCR conversion from RNA to DNA), normally only one strand (forward or reverse) of the DNA is actually sequenced. In this case, the researcher must decide whether the output signal, and the resulting basecall is accurate based on their experience and skill in reading sequence signals. If the signal and resulting basecall is of questionable reliability, then the researcher must start the sequencing run again in the hope of obtaining a better signal.

In some cases, the forward and reverse stands are both sequenced, such as by using two dyes on a MICROGENE CLIPPER sequencer manufactured by Visible Genetics Inc. Forward and Reverse strand sequencing provides the researcher with more information and allows the researcher to evaluate the quality and reliability of the data from both strands. If the bases on both strands complement each other as expected, then this helps to confirm the reliability of the sequence information. However, in some instances, after the signal data from sequencing is assigned a base (e.g. A, C, G or T), the corresponding base on the opposite strand does not match. If the signal and resulting basecall is of questionable reliability, then the researcher must start the sequencing run again in the hope of obtaining better signal. Alternatively, the researcher might manually review ("eyeball" analysis) the signal data from both the forward and reverse strands and make a decision on which strand's data was more reliable. Unfortunately, any such decision will vary between individual researchers and can lead to inconsistent determination of reliablity within the same sequencing run. Furthermore, this kind of eyeball analysis requires special training which makes it poorly suited for application in routine diagnostic applications.

It would therefore be desirable to have a method for sequencing nucleic acid polymers in which discrepancies can be resolved using automated procedures, i.e. using computerized data analysis. It is an object of the present invention to provide such a method, and an apparatus for performing the method.

Summary of the Invention

In accordance with the invention, nucleic acid polymers are sequenced in a method

comprising the steps of

(a) obtaining forward and reverse data sets for forward and reverse strands of the sample nucleic acid;

(b) determining the apparent sequence of bases for the forward and reverse data sets;

(c) comparing the apparent forward and reverse sequences of basis for perfect complementarity to identify any deviations from complementarity in the apparent sequence, any such deviation presenting a choice between two bases, only one of which is correct;

(d) applying a confidence algorithm to peaks in the data set associated with a deviation to arrive at a numerical confidence value; and

(e) comparing each numerical confidence value to a predetermined threshold and selecting as the correct base the base represented by the peak which has the better numerical confidence value, provided that the numerical confidence value is better than the threshold.

The confidence algorithm takes into account at least one, and preferably more than one of several specific characteristics of the peaks in the data sets that were not complimentary.

Brief description of the drawing

The invention will be described with respect to a drawing in several figures, of which:

Fig. 1 shows four regions of the HIV-1 genome sequenced in the analysis of HIV according to the invention;

Fig. 2 shows the improvement in accuracy in selecting one of two HIV species using both

forward and reverse strands;

Figs. 3 and 4 show a comparison of text files representing apparent base sequences;

Fig. 5 shows a schematic representation of the method of the invention in a case where a deviation between forward and reverse sequences is observed; and

Fig. 6 shows sequence data for forward and reverse strands in which regularity/evenness of peak separation can be used as a key characteristic in determining a numerical confidence value.

Detailed Description of the Invention

The purpose of the present invention is to provide a novel method and system for the reduction of errors in sequencing data, and in particular to provide a method and system which can automate the process of reconciling forward and reverse strand sequences to readily provide sequencing results of improved quality.

In the present disclosure, the invention is illustrated using sequence data taken from the TruGene HIV-1 Assay manufactured by Visible Genetics Inc. In this case, data traces containing sequence information for one amplicon from the Protease region and three amplicons from the reverse transcriptase (RT) region as shown in Figure 1 were considered. The reference to this sequence is provided for purposes of example only, however, and to demonstrate the efficacy of the invention. Thus, in a broader sense, the present invention may be applied to the sequencing and error correction of sequencing data for any polynucleotide, including DNA and RNA sequences for any gene or gene fragment.

Error rates in HIV mutation sequencing are in the range of 5 errors/1000 bases sequenced or higher for many homebrew sequencing methods (single strand). Using the method of the invention these rates are substantially reduced to provide error rates that routinely are as low as 5 errors/100,000 bases and may reach levels as low as 2.5/1000000 bases for a 300 base call. Figure 2 shows the improvement in accuracy in detecting one of two HIV species using both forward and reverse strands.

The method of determining the sequence of a sample polynucleotide in accordance with the invention involves the following basic steps:

(a) obtaining forward and reverse data sets for the sample polynucleotide;

(b) identifying the sequence of bases within the forward and reverse data sets;

(c) comparing the sequence of bases within the forward and reverse data sets to identify any deviations from perfect complementarity in the sequences as determined for the two sets; and

(d) applying a confidence algorithm to each deviation to select the correct base from between the choices presented by the identified forward and reverse sequence.

A variety of procedures for obtaining the forward and reverse data sets for the sample polynucleotide are known, and all can be applied in the present invention. In general, the sample polynucleotide or a complementary copy of the sample polynucleotide is combined with a sequencing primer which is extended using a template-dependent polymerase enzyme in the presence of a chain-terminating nucleotide triphosphate (e.g. a dideoxynucleotide) to produce a set of sequencing fragments the lengths of which reflect the positions of the base corresponding to the dideoxynucleotide triphosphate in the extended primer. By preparing one set of fragments for each type of base (e.g. A, C, G and T), the complete sequence for the sample polynucleotide is determined. Forward and reverse sequences are obtained by utilizing two primers which hybridize to the two strands of a duplex DNA molecule.

The preparation of fragments mixtures providing forward and reverse sequencing data sets can be performed as individual reactions, or it can be concurrent. In a concurrent procedure, forward and reverse primers with different labels are extended in the same reaction mixture. This process can involve a single extension cycle as disclosed by Wiemann et al., *Anal. Biochem* 224: 117-121 (1995), or multiple bi-directional cycles (preferably using CLIP™ sequencing chemistry, Visible Genetics Inc.) as described in International Patent Publication No. WO 97-41259 entitled "Method for sequencing of nucleic acid polymers" each of which are incorporated herein by reference. The process can also involve multiple bi-directional cycles as described in US Patent Application Serial No. 09/009,483, incorporated herein by reference to the extent permitted. Thus, fragment mixtures reflecting the sequence of the forward and reverse strands of the same

polynucleotide are obtained by multiple cycles of a primer extension reaction in which two differently and distinguishably labeled primers are extended in the presence of chain terminator nucleotides in a single reaction mixture. Preferred fragment mixtures utilize fluorescent labels which are detected following electrophoretic separation to produce a forward and reverse data trace for each base position.

The next step in the method of the invention is the identification of the apparent base sequence for both the forward and reverse strands of the sample polynucleotide, a process sometimes referred to as "base-calling." The process of base-calling is theoretically quite straightforward, requiring nothing more than the sequential reading of the bases from the overlapping data traces to produce a list of bases reflecting the sequence. In practice, the process is more complicated, because of deparatures of actual data from the theoretical ideal. As for the initial generation of the data traces, there are various methods known for dealing with these complications to facilitate automated base-calling from real data, including those disclosed in US Patents Nos. 5,365,455 and 5,502,773, which are incorporated herein by reference.

A preferred base-calling technique is that disclosed in US Patent No. 5,853,979 entitled "Method and system for DNA sequence determination and mutation detection with reference to a standard" and International Patent Publication WO 97-02488 entitled "Method and system for DNA sequence determination and mutation detection," each of which are incorporated herein by reference. In this method, a fragment pattern representing the positions of a selected nucleic acid base within the polymer as a function of migration time or distance is evaluated to determine one or more "normalization coefficients." These normalization coefficients reflect the displacement, stretching or shrinking, and rate of stretching or shrinking of the fragment pattern, or segments thereof, which are necessary to obtain a suitably high degree of correlation between the fragment pattern and a standard fragment pattern which represents the positions of the selected nucleic acid base within a standard polymer actually having the known sequence as a function of migration time or distance. The normalization coefficients are then applied to the fragment pattern to produce a normalized fragment pattern which is used for base-calling in a conventional manner.

The process of comparing the experimental fragment pattern and the standard

fragment pattern to arrive at normalization coefficients can be carried out in any number of ways without departing from the present invention. In general, suitable processes involve consideration of a number of trial normalizations, and selection of the trial normalization which achieves the best fit in the model being employed. It will be understood, however, that the theoretical goal of achieving an exact overlap between an experimental fragment pattern and a standard fragment pattern may not be realistically achievable in practice, nor are repetitive and time consuming calculations to obtain perfect normalization necessary to the successful use of the invention. Thus, when employing this method to facilitate base-calling, the term "high degree of normalization" refers to the maximization of the normalization which is achievable within practical constraints. As a general rule, a point-for-point correlation coefficient calculated for normalized fragment patterns and the corresponding standard fragment pattern of at least 0.8 is desirable, while a correlation coefficient of at least 0.95 is preferred.

The result of the base-calling is two lists of bases, one for the forward strand and one for the reverse strand. Each list contains an indication of the base at a particular location (e.g. one-letter indications A, C, G and T). In some instances, a list may include one or more blanks. Such blanks are inserted by the alignment program to maximize the extent of alignment and take into account the fact that insertions or deletions within one strand may result in a shift of one portion of the strand relative to the corresponding portion of the other strand. These two lists are suitably stored in a data processor performing the sequence analysis as text files. The next step is the comparison of these two text files to determine whether there are any deviations from the theoretically expected perfect complementarity. This comparison process can be performed by any of several methods. Common to these methods is the appropriate alignment of the text listings of bases to a common starting point. This alignment involves an iterative testing of various alignment options to arrive at the best alignment. Iterative routines for accomplishing this alignment have been disclosed by Needleman et al., "A general method applicable to the search for similarities in amino acid sequences of two proteins" *J. Mol. Biol.* 48: 443-453 (1970) and Smith et al., "The identification of common molecular subsequences" *J. Mol. Biol.* 147: 195-197 (1981).

In a first approach, the text file listings of the forward strand and the reverse strand are each aligned with a text file listing of a standard sequence for the sample

polynucleotide being sequenced (eg. HIV-1 wild-type sub-type B in the case of the HIV example discussed below). Alternatively, the text file listings for the forward and reverse strands can be aligned to each other. It will be appreciated that the use of text files is only one option, however, and that the alignment may occur between the experimental data sets, or between the experimental data sets and a reference data set.

The important aspect is that the alignment step produce information which will allow a determination of whether or not there is a deviation in the sequence of the forward and reverse strands from the expected complementarity. When such a deviation is detected, the method of the invention provides an automated system for selecting between the options presented and generating a "correct" sequence. This selection process can take place in several steps using a confidence algorithm.

The confidence algorithm is used to assign a confidence value to each base in the forward or reverse text listing that is not confirmed by the other listing. The confidence value is a measure of the likelihood that a particular base identified in a text listing is the correct base. The confidence algorithm determines the confidence value for a peak by taking into account a variety of factors which reflect the quality of the data traces. Specific factors include:

1. separation distance between peaks;
2. regularity/evenness of peak separation;
3. peak height compared to neighbors (higher confidence if similar);
4. peak area compared to neighbors (higher confidence if similar);
5. distance to neighbors compared to the local average distance to neighbors;
6. resolution of the peak (lower confidence for lower resolution); and
7. signal-to-noise ratio in the region around the peak (lower confidence as the peak's size is more similar to the noise level).

The number of characteristics and the particular characteristics considered are a matter of design choice which is driven by the performance of the combination of chemistry and instrumentation which is used. In some systems, it may be the case that a few characteristics (e.g. two) are particularly sensitive to the causes of error, in which case determination of a numerical confidence value based on these characteristics is sufficient.

In a preferred embodiment, all of these factors are included in a weighted

combination to arrive at the confidence value, although the use of less than all of the factors may be considered, particularly where two factors are similar (such as peak height and peak area). The confidence value is also lowered in some recognized special cases:

- peak is a heterozygote;
- more than two overlapping peaks; or
- the peak is small compared to its neighbors.

The system evaluating the data traces may also attempt to fit groups of peaks to the signal when the peaks are low resolution. These fitted peaks are also assigned confidence values using the above.

The specific weighting applied to each factor will vary to some extent with the configuration of the sequencing instrument employed and the chemistry used, since each experimental combination will tend to produce different variability which effects the accuracy of the base call. Thus, for example, some sequencing chemistries are prone to greater variability in peak height than others, such that variations in peak height might be of less significance in the confidence algorithm. Initial determination of the appropriate weighting for a given system can be done using multiple calibration runs with a known sequence and varying the weight given to the different factors to arrive at the most consistent and error-free results. In a preferred embodiment, the weights given to the different factors are updated heuristically as experimental sequences are determined. The attached source code provides this feature.

Once the weights to be given to the confidence factors are determined, an overall numerical confidence value is calculated for each peak which indicates deviations from the expected match between the forward and reverse sequences. This calculated confidence value is then compared to a predetermined threshold value to determine whether the confidence value is sufficiently high (assuming that the characteristics are combined such that a larger number is indicative of high confidence) to accept the base as being correct. It will be appreciated that the numerical value of this threshold will depend on many factors, including the units of the measurements used for the individual factors and the level of rigor which the individual user of the invention chooses to apply. Thus, it is not possible to give meaningful numerical examples of a threshold value. Preferably the threshold value should, however, be one which when applied in combination with the selected weights for the various

factors to standard sequence (such as M13) produces error rates of less than 1/1,000 bases over the first 300 bases of the region sequenced.

It will be appreciated that the form of the confidence algorithm can be manipulated such that a "good" result will be either high or low. When the algorithm is such that a "good" result is large, then the numerical confidence value is "better" then the threshold if it exceeds the threshold. When the algorithm is such that a "good" result is small, then the numerical confidence value is better if it is less than the threshold. Similarly, in comparing the two numerical confidence values, the one that is "better" is one that is larger in the first circumstance and smaller in the second.

The application of the confidence values to the actual sequences is suitably performed in several successive steps. If the forward and reverse sequences do not confirm each other, then if a base exists (as opposed to a blank) in both the forward and reverse experimental sequences and the confidence measure of the better of the two is above the confidence threshold currently set, then that base with the higher confidence measure is assigned. If both bases are above the confidence threshold, an additional comparison to the reference sequence may be carried out, with the base which is the same as the reference being selected in this instance. Identity with the reference sequence is not a basis for selection as the correct base in an experimental sequence in the absence of a sufficient confidence value.

If only one of the experimental sequences has a base at the location of the deviation and there is a base (as opposed to a blank) in the reference sequence and the identified base in the experimental sequence is above the confidence threshold, then the base from the experimental sequenceis used as the "correct" base (the base type in the reference sequence is ignored, only the spacing information is used). If there is a blank in the reference sequence and that is confirmed by either of the experimental sequences, then a blank should be put in the corrected sequence. If none of the above conditions apply, it is recommended to put an 'N' in the output sequence (standard letter denoting all bases are present) and mark the location as uncorrected.

The method of the invention is preferably carried out in an apparatus or system running appropriate computer code. The apparatus or system comprises at least a data processor operably programmed to perform the steps of identifying the sequence of bases within the forward and reverse data sets; comparing the sequence of bases within the forward

and reverse data sets to identify any deviations from perfect complementarity in the sequences as determined for the two sets; and applying a confidence algorithm to each deviation to select the correct base from between the choices presented by the identified forward and reverse sequence. The apparatus or system further comprises means for obtaining forward and reverse data sets for the sample polynucleotide. In the case of an integrated system, this may be a direct data feed from an electrophoresis apparatus connected to the data processor. In a distributed system, the data sets can be obtained via a connection on a local area network (LAN), a wide area network (WAN), by modem or cable modem transmission or by insertion of a portable storage medium (diskette, tape etc.) into a drive capable of reading the portable storage medium. The apparatus or system further comprises means for providing useful output of the determined sequence. This may be as a video display or as a sequence listing stored on a storage medium such as a disk drive or read/write CD-ROM.

Figures 3 and 4 show screen output from an Intel® processor-based Hewlett-Packard Vectra VL computer (running an OpenStep Mach operating system) in which one amplicon of the RT region is being corrected for sequencing errors by analyzing both strands (shown as 3 prime (text 1) and 5 prime (text 2)). The highlighted bases on the reference show places of disagreement between the two strands (text 1 and text 2). The highlighted text in the Corrected area represent corrections according to the above embodiment of the present invention. In this case, the software allows adjustment of the confidence threshold, which as shown is set to 80%.

Figure 5 shows a chart indicating possible outcomes if basecall for a particular base is not confirmed by each strand. In this case, if the base in text 1 and text 2 are a, then they have confirming sequence (e.g. forward strand is A and reverse strand is T). N/A means that the confidence value for a particular base for either strand was below threshold (50% in this figure) and correction was not possible.

Figure 6 shows sequence data for the forward and reverse strand in which the parameter of "regularity/evenness of peak separation" has been measured as lane-to-lane shift in seconds (y-axis) with the base location on the x-axis. The reverse strand shows an erratic shift in lane timing and indicates a compression in signals which is unreliable. In contrast, this parameter shows more regular peak separation in the forward strand which may provide

for an overall higher confidence value for bases on the forward strand.

In the course of evaluating the method and system of the invention with HIV sequence data, it has been observed that errors in the sequence data for one strand tend to be random with respect to errors in the sequence data of the opposite strand. As a result, it is less likely that an error will occur in the sequence data of both strands at the same base position. These errors can thereby be corrected should the reliability of one strand (confidence) be of a sufficient level. This observation highlights the utility of the present invention to create corrected sequence data in HIV genotyping, and it is believed that comparable benefits will be obtained for sequencing in general.

What is claimed is:

1. A method for determining the sequence of a sample nucleic acid polymer comprising the steps of:

(a) obtaining forward and reverse data sets for forward and reverse strands of the sample nucleic acid;

(b) determining the apparent sequence of bases for the forward and reverse data sets;

(c) comparing the apparent forward and reverse sequences of basis for perfect complementarity to identify any deviations from complementarity in the apparent sequence, any such deviation presenting a choice between two bases, only one of which is correct;

(d) applying a confidence algorithm to peaks in the data set associated with a deviation to arrive at a numerical confidence value; and

(e) comparing each numerical confidence value to a predetermined threshold and selecting as the correct base the base represented by the peak which has the better numerical confidence value, provided that the numerical confidence value is better than the threshold.

2. The method of claim 1, wherein the numerical confidence value results from a selected combination of two or more selected characteristics of each peak associated with a domain name.

3. The method of claim 2 wherein the characteristics are selected from among the following: separation distance between peaks; regularity/evenness of peak separation; peak height compared to neighbors (higher confidence if similar); peak area compared to neighbors (higher confidence if similar); distance to neighbors compared to the local average distance

to neighbors; resolution of the peak (lower confidence for lower resolution); signal-to-noise ratio in the region around the peak (lower confidence as the peak's size is more similar to the noise level).

4. The method of claim 2 or 3, wherein the numerical confidence value is a weighted combination of the selected characteristics.

5. The method of claim 4 wherein weights applied to each of the selected characteristics are initially determined for a combination of chemistry and instrument from a plurality of calibration runs performed using the given combination of chemistry and instrumentation.

6. The method of claims 4 or 5, wherein weights applied to each of the selected characteristics are updated based upon accumulated data obtained when sequencing sample nucleic acid polymer.

7. The method of any of claims 1-6, wherein the numerical confidence results from a combination of at least the following characteristics of each peak associated with a deviation:

separation distance between peaks;
regularity/evenness of peak separation;
peak height compared to neighbors;
peak area compared to neighbors;
distance to neighbors compared to the local average distance to neighbors.
resolution of the peak; and
signal-to-noise ratio in the region around the peak.

8. An apparatus for sequencing a sample nucleic acid polymer in accordance with the method of claims 1-7.

9. A method for determining the sequence of a sample polynucleotide comprising the steps of:

(a) obtaining forward and reverse data sets for the forward and reverse strands of the sample polynucleotide by multiple cycles of a primer extension reaction in which two labeled primers are extended in the presence of chain terminators nucleotides in a single reaction mixture;

(b) determining the apparent sequence of bases for the forward and reverse data sets;

(c) comparing the apparent forward and reverse sequences of basis for perfect complementarity to identify any deviations from complementarity in the apparent sequence, any such deviation presenting a choice between two bases, only one of which is correct;

(d) applying a confidence algorithm to peaks in the data set associated with a deviation to arrive at a numerical confidence value; and

(e) comparing each numerical confidence value to a predetermined threshold and selecting as the correct base the base represented by the peak which has the better numerical confidence value, provided that the numerical confidence value is better than the threshold.

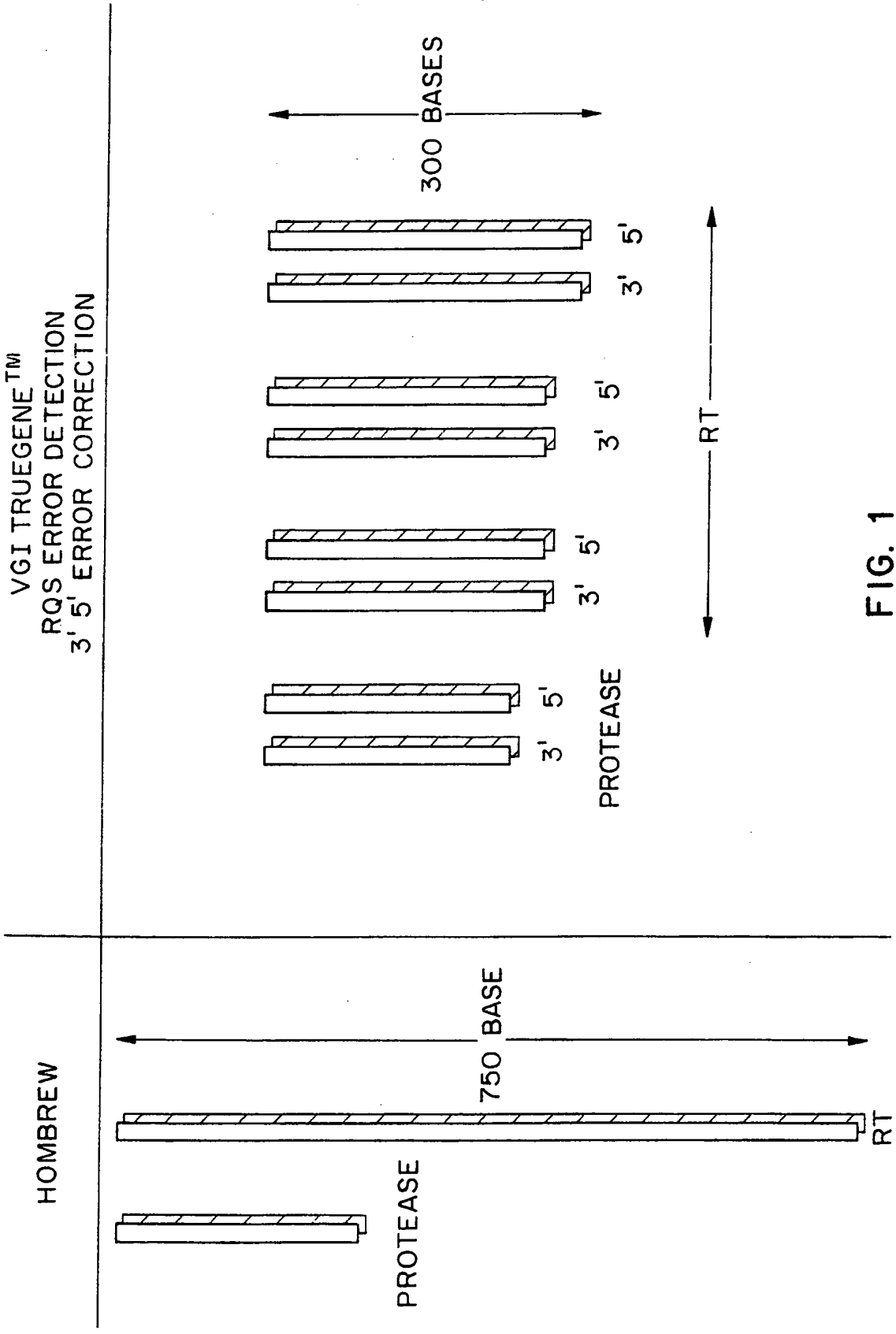

VGI TRUEGENE™
RQS ERROR DETECTION
3' 5' ERROR CORRECTION
300 BASES
3' 5'
3' 5'
3' 5'
3' 5'
3' 5'
RT
PROTEASE
HOMBREW
750 BASE
RT
PROTEASE

FIG. 1

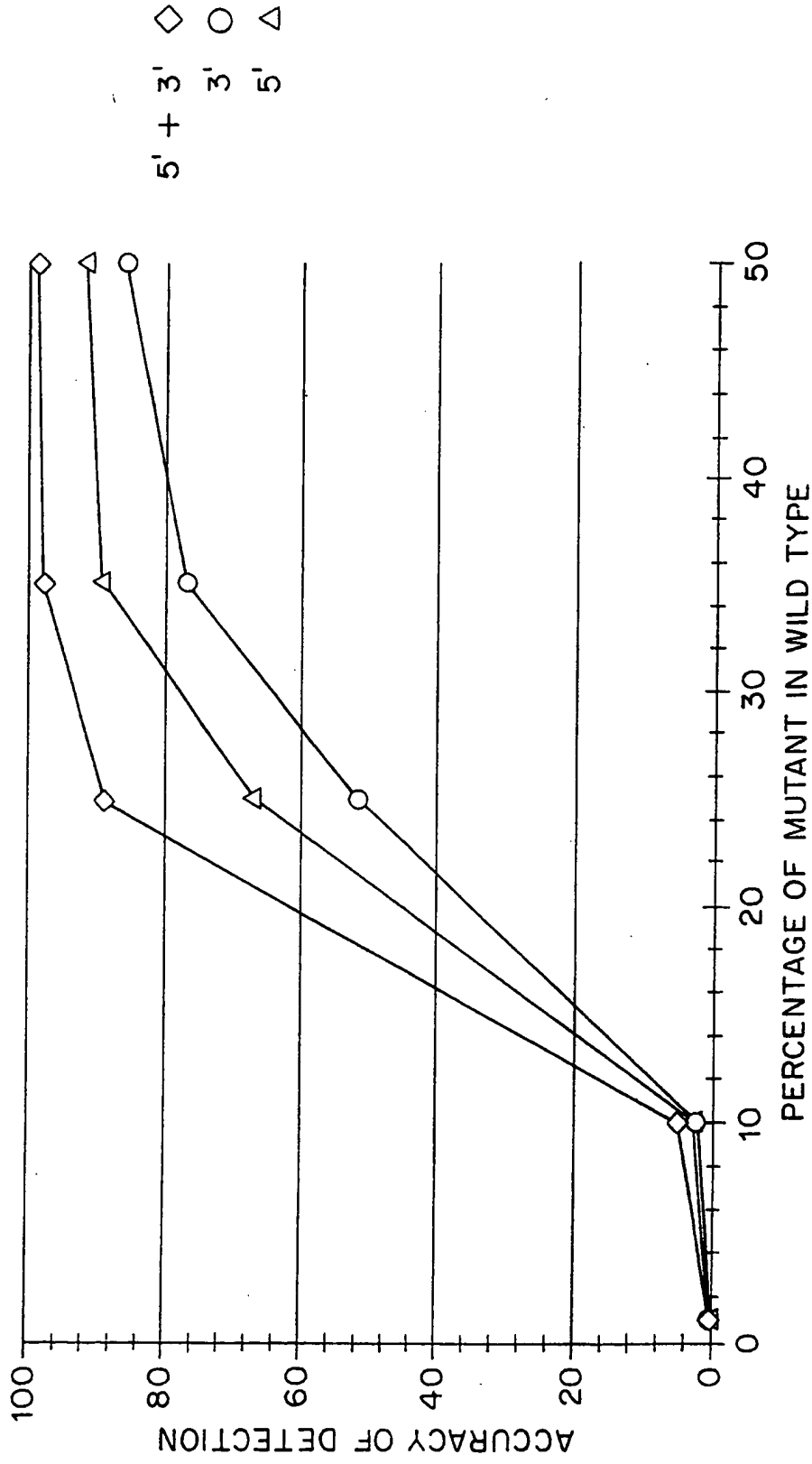

5' + 3'
3'
5'
ACCURACY OF DETECTION
100
80
60
40
20
0
0
10
20
30
40
50
PERCENTAGE OF MUTANT IN WILD TYPE

FIG. 2

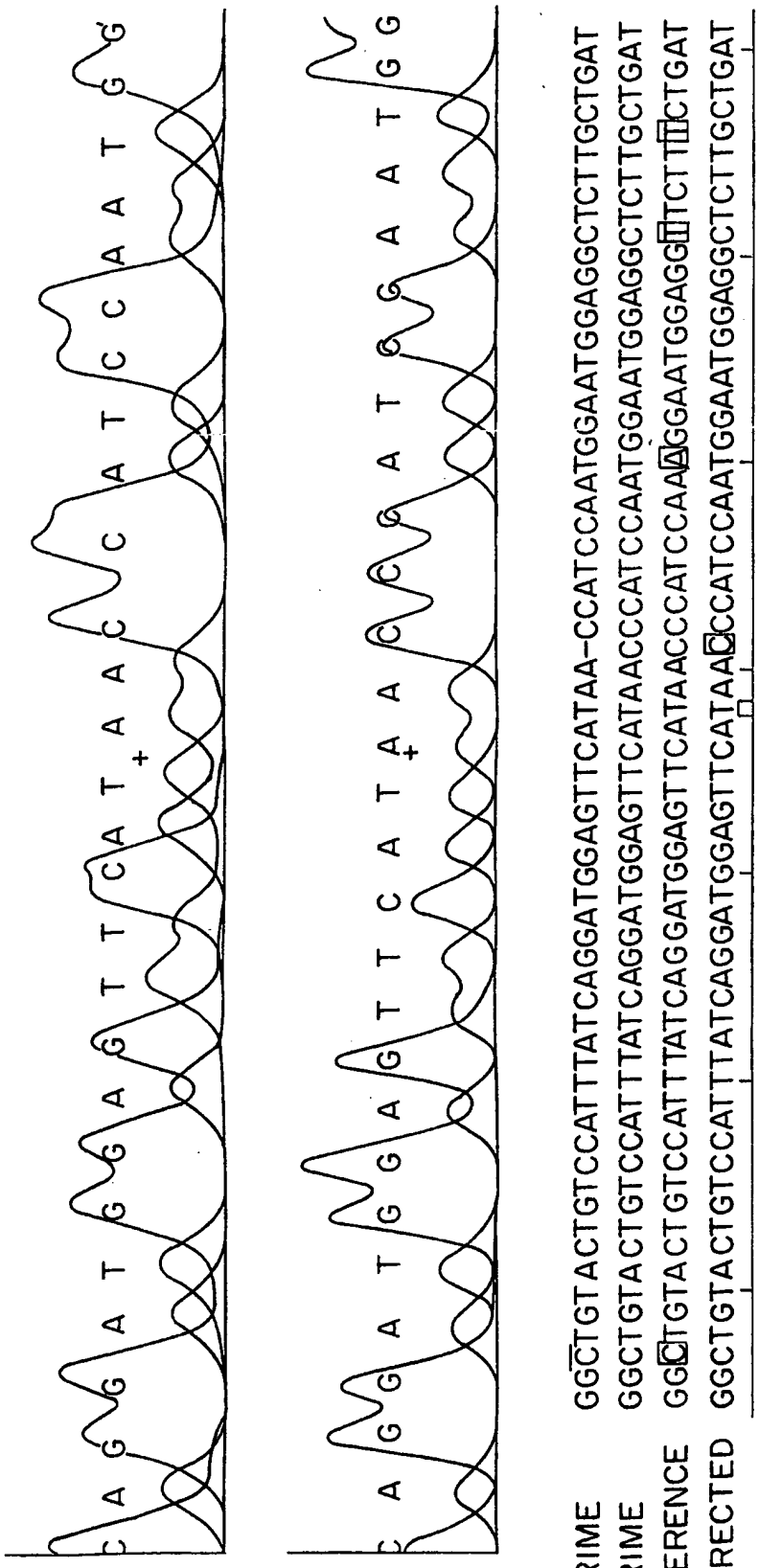
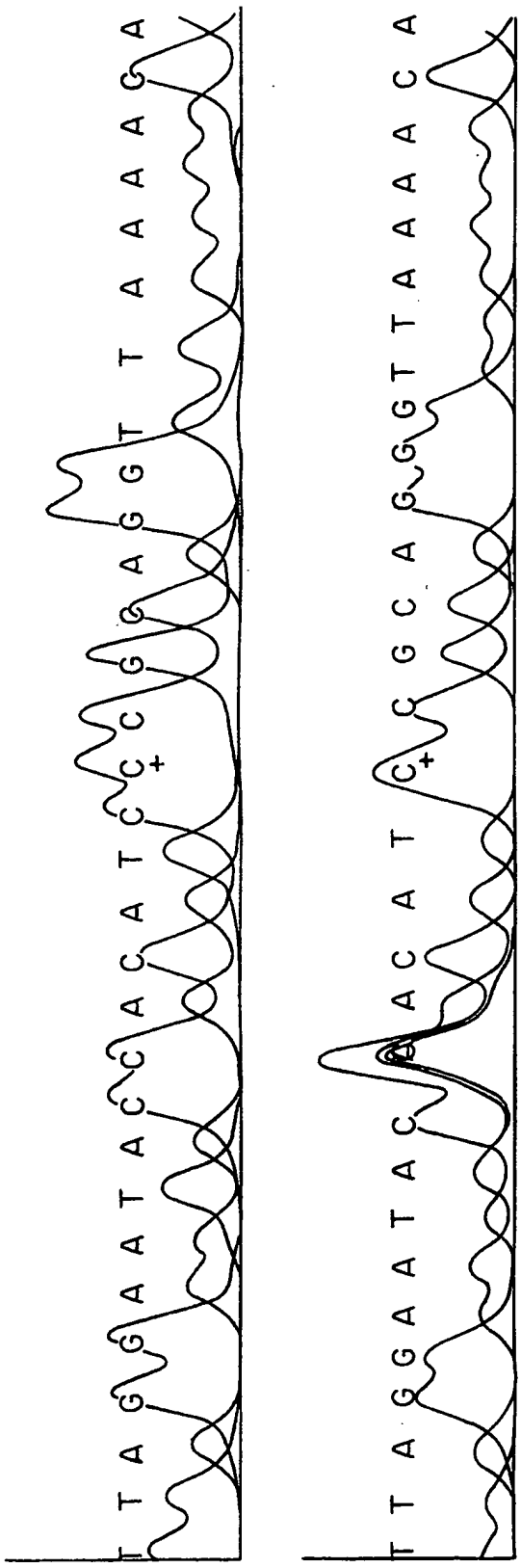

3 PRIME GGCTGTACTGTCCATTTATCAGGATGGAGTTCATAA-CCATCCAATGGAATGGAGGCTCTTGCTGAT
5 PRIME GGCTGTACTGTCCATTTATCAGGATGGAGTTCATAACCCATCCAATGGAATGGAGGCTCTTGCTGAT
REFERENCE GGCTGTACTGTCCATTTATCAGGATGGAGTTCATAACCCATCCAAGGAATGGAGGCTCTTCTGAT
CORRECTED GGCTGTACTGTCCATTTATCAGGATGGAGTTCATAACCCATCCAATGGAATGGAGGCTCTTGCTGAT

FIG. 3



3 PRIME CTGCTCTGGGAAGTTCAATTAGGAATACCACATCCCGCA-GGTTAAAACAGAAGAAATCAGTAACAG
5 PRIME CT--TCTGGGAAGTTCAATTAGGAATACCAACATCCGCAGGGTTAAAACAGAAGAAATCAGTAACAG
REFERENCE CT--TCTGGGAAGTTCAATTAGGAATACCACATCCCGCAGGGTTAAAACAGAAGAAATCAGTAACAG
CORRECTED CT□□TCTGGGAAGTTCAATTAGGAATACCACACCCGCAGGGTTAAAACAGAAGAAATCAGTAACAG

FIG. 4

RESULT	TEXT 1	TEXT 2	REFERENCE TEXT	CONFIDENCE VALUE * SIGNAL 1	CONFIDENCE VALUE * SIGNAL 2
a	a	a	b		
a	a	b	a	80%	60%
a	a	b	c	80%	60%
a	a	—	b	80%	—
h/a	a	—	b	40%	—

* CONFIDENCE THRESHOLD = 50%

FIG. 5

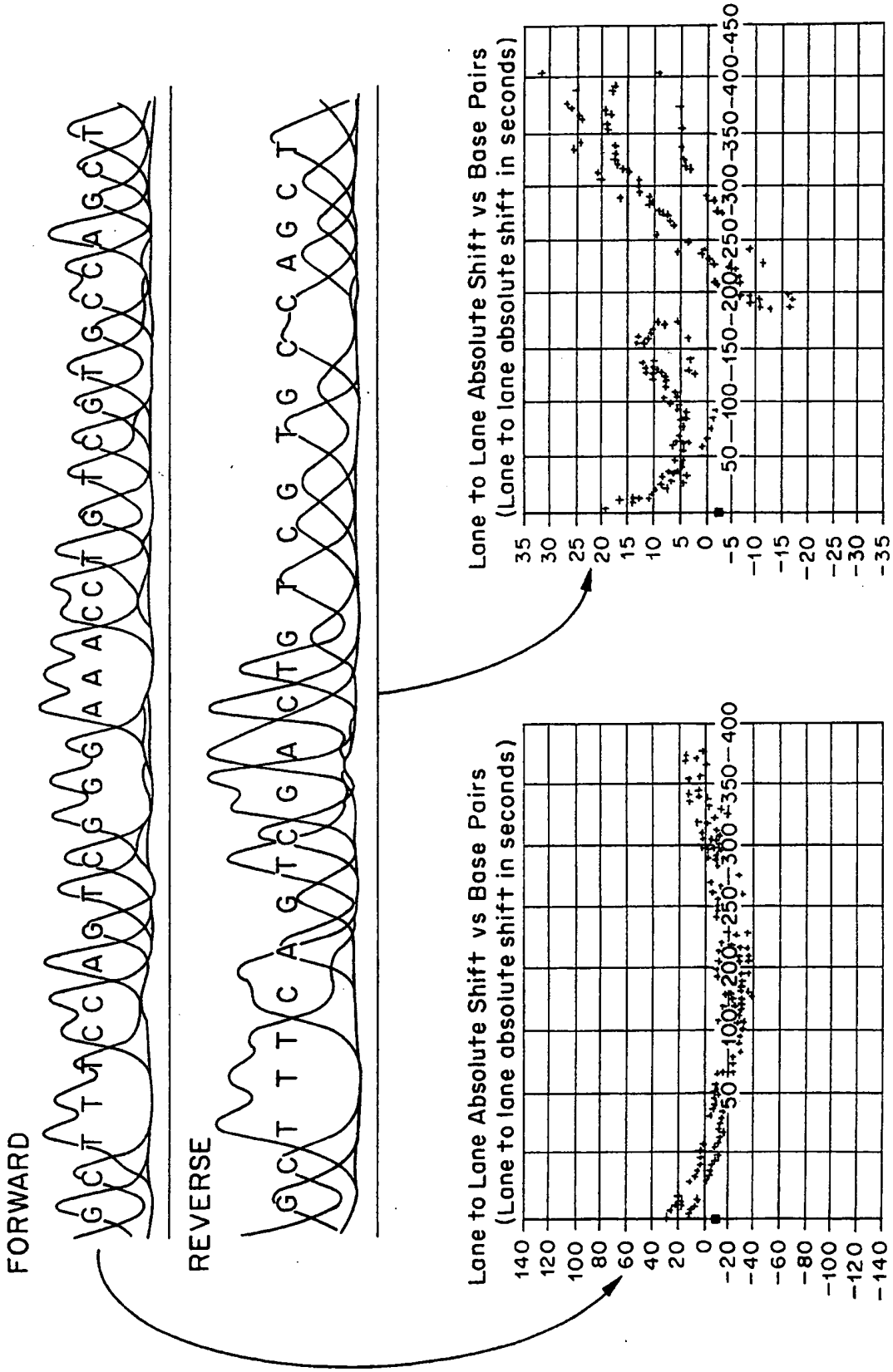

FORWARD
REVERSE
Lane to Lane Absolute Shift vs Base Pairs
(Lane to lane absolute shift in seconds)
Lane to Lane Absolute Shift vs Base Pairs
(Lane to lane absolute shift in seconds)

FIG. 6